DOCUMENTO

IA GENERATIVA

EL CONSUMO ENERGÉTICO DE LA IA GENERATIVA

OCTUBRE DE 2024

Ametic
LA VOZ DE LA INDUSTRIA DIGITAL

Contenido

El consumo energético de la IA generativa

La inteligencia artificial generativa ofrece innumerables aplicaciones y usos que, hasta hace poco, parecían inalcanzables. No obstante, esta tecnología emergente enfrenta diversos desafíos y limitaciones que deberán ser superados en el futuro cercano. Uno de los principales retos es el alto consumo energético requerido, tanto en la fase de entrenamiento como durante la ejecución de los modelos.

Estudios recientes[1] señalan que, si la tendencia no cambia, la IA podría estar en camino de consumir anualmente tanta electricidad como todo el país de Irlanda (29,3 Tera vatios-hora por año). Este alto consumo, además, comporta la generación de CO_2. Por ejemplo, el entrenamiento de Llama 3.1[2] en sus tres versiones (8B, 70B y 405B) ha generado 11390 toneladas de CO_2.

¿Por qué la IA generativa consume tantos recursos?

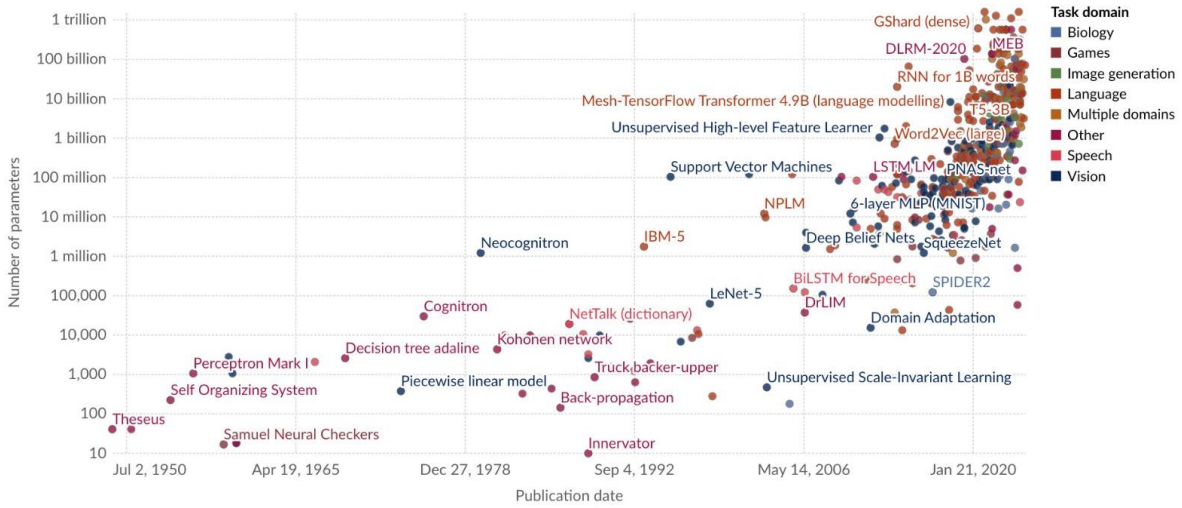
La IA generativa consume enormes cantidades de recursos computacionales debido a varios factores, entre ellos el tamaño y complejidad de los modelos, la cantidad de datos requeridos para el entrenamiento y los cálculos involucrados en el entrenamiento y ejecución.

Número de parámetros

El crecimiento exponencial en el número de parámetros en los modelos es uno de los principales motivos de este alto consumo de recursos. Los parámetros, como mencionamos anteriormente, son los pesos y sesgos que se ajustan durante el entrenamiento para que el modelo aprenda a generar contenido útil. Cuantos más parámetros tenga un modelo, más complejo y "capaz" es, ya que tiene una mayor capacidad para aprender patrones detallados y representaciones complejas a partir de los datos.

El número de parámetros en los modelos de IA ha crecido exponencialmente en los últimos años. Por ejemplo, el modelo GPT-2 de OpenAI (2019) tenía alrededor de 1.5 mil millones de parámetros, mientras que GPT-3 (2020) aumentó a 175 mil millones, y se estima (no se han publicado los datos) que GPT-4 tiene 1.76 billones de parámetros (billones de los españoles, no anglosajones😊). Otros modelos similares open source, como Llama o Mistral, siguen esta tendencia.

[1] https://www.cell.com/joule/abstract/S2542-4351(23)00365-3
[2] https://github.com/meta-llama/llama-models/blob/main/models/llama3_1/MODEL_CARD.md



Los parámetros son variables en un sistema de IA cuyos valores se ajustan durante el entrenamiento para determinar cómo los datos de entrada se transforman en la salida deseada. Source: https://ourworldindata.org/

Este aumento masivo en la cantidad de parámetros requiere una mayor capacidad de procesamiento y memoria, ya que cada cálculo involucra cientos de miles de millones de operaciones matemáticas. El número exacto de operaciones estimadas requeridas para entrenar un modelo de última generación, también llamado modelo frontera, es de $1e^{25}$.

En este punto vale la pena mencionar que empíricamente se ha observado una tendencia conocida como “rendimientos decrecientes” (diminishing returns, en inglés). Aunque el rendimiento de los modelos mejora al aumentar su tamaño, las ganancias se vuelven cada vez más marginales. Por esta razón los modelos de IA generativa basados en transformers han crecido exponencialmente en el número de parámetros mientras sus capacidades solo han aumentado de forma lineal. Por ejemplo, un estudio de OpenAI (“Scaling Laws for Neural Language Models”)[3] muestra que la mejora en la “pérdida” del modelo (una medida de su error) sigue una relación de ley de potencia con respecto al tamaño del modelo, el tamaño del conjunto de datos y la cantidad de computación utilizada para el entrenamiento. Sin embargo, las mejoras adicionales se vuelven más pequeñas a medida que se incrementan los parámetros, especialmente cuando se superan ciertos umbrales. Esto sugiere que, aunque aumentar el tamaño de los modelos resulta en un mejor rendimiento, el ritmo de mejora no es exponencial y se acerca más a un crecimiento lineal o sublineal.

[3] https://openai.com/index/scaling-laws-for-neural-language-models/

Datos de entrenamiento

Para que un modelo generativo funcione eficazmente, debe entrenarse con datos como texto, imágenes o vídeo. El proceso de entrenamiento implica ajustar los miles de millones de parámetros a lo largo de muchas iteraciones por la red neuronal. Esto requiere no solo de gran cantidad de procesamiento paralelo (usualmente en GPUs), sino también de almacenamiento de memoria y de ancho de banda para manejar los datos de entrenamiento y los estados intermedios del modelo. Además, las computadoras deben realizar operaciones matemáticas como multiplicaciones de matrices, de manera eficiente y rápida.

Para hacernos una idea de la cantidad de datos usados, un modelo como Llama 3 de 70 mil millones de parámetros (Llama 3 70B), liberado en abril de 2024, ha sido entrenado con 15 billones de tokens* (billones de los españoles)[4]. Si lo comparamos con otros datasets similares[5], esto serían aproximadamente 44 Terabytes de datos.

> * *Un **token** es una unidad básica de texto que los modelos de lenguaje utilizan para procesar datos. Puede ser una palabra completa, una parte de una palabra, o incluso un solo carácter, dependiendo de cómo el modelo divida el texto.*

Complejidad de la arquitectura hardware y software

Por otro lado, el crecimiento exponencial en el número de parámetros, y el mecanismo de atención utilizado en los *transformers,* que hace que cada palabra o elemento en la entrada esté vinculado a cada otro (lo que significa que los cálculos aumentan cuadráticamente en relación con el tamaño de la secuencia), implica una mayor necesidad de recursos hardware (memoria RAM, almacenamiento, comunicaciones, etc.).

En un post reciente, Meta publica[6] que se han necesitado dos clústeres de 24.000 GPUs para entrenar Llama 3 (abril 2024), y que para finales de 2024 esperan contar con un clúster de 350.000 GPUs. Y aunque este incremento en el número de GPUs pudiera parecer exagerado, la organización sin ánimo de lucro Epoch AI, publica[7] que la capacidad de cómputo para entrenar modelos frontera se multiplica por 4 o 5 cada año.

[4] https://ai.meta.com/blog/meta-llama-3/

[5] https://huggingface.co/datasets/HuggingFaceFW/fineweb

[6] https://engineering.fb.com/2024/03/12/data-center-engineering/building-metas-genai-infrastructure/

[7] https://epochai.org/blog/training-compute-of-frontier-ai-models-grows-by-4-5x-per-year

Si revisamos el 'model card'[8] (ficha del modelo), de Llama 3.1 (julio 2024) de Meta, veremos que se han requerido de 39,3 millones de horas de procesamiento, en el hardware más avanzado de Nvidia, las GPU H100-80GB.

Impacto energético y medioambiental

La IA generativa tiene un **impacto energético** debido a los motivos expuestos en el apartado anterior. Tomemos como ejemplo los modelos de Llama 3.1 en sus diferentes versiones (8B, 70B y 405B). Para entrenar estos modelos, se utilizaron **39,3 millones de horas de GPU**. Este nivel de uso Meta lo traduce en una emisión de **11.390 toneladas de CO2**. Para poner esto en perspectiva, esa cantidad de emisiones es similar a las emisiones anuales de más de 2.400 vehículos de pasajeros o la energía utilizada por 1.000 hogares en un año. Aunque dada la reciente noticia[9] de The Guardian respecto a los grandes hyperscalers infra-reportando sus emisiones de CO2, esta cifra podría ser mucho más alta.

En un reciente artículo[10] de Epoch AI se analiza la viabilidad de escalar los modelos frontera al ritmo actual, y se proyecta que en 2030 el principal factor limitante va a ser la disponibilidad de energía para el entrenamiento. Aun suponiendo que estos futuros modelos se entrenaran en hardware 24 veces más energéticamente eficientes que el hardware actual, esta es la mejora conseguida hasta ahora, aun se requeriría 200x más energía de la usada por Llama 3.1, unos 6 GW de potencia. Por dar un orden de magnitud, España tiene instalada aproximadamente 7 GW de potencia nuclear.

Aunque hay menos información disponible, la ejecución de modelos (inferencia) también tiene un alto impacto. Algunas estimaciones[11] sugieren que en enero de 2023 OpenAI usaba 30.000 GPUs para manejar millones de peticiones diarias de sus usuarios, y que esas peticiones consumían alrededor de 1GWh cada día[12]. El equivalente a 33.000 hogares.

Algunos análisis[13] apuntan que en 2030 los centros de datos para IA consumirán el 4,5% de la energía global generada. El consumo de los centros de datos pasará de 49 GW en 2023 a 96 GW en 2026, de los cuales 40GW será para IA. Este mismo análisis también

[8] https://github.com/meta-llama/llama-models/blob/main/models/llama3/MODEL_CARD.md
[9] https://www.theguardian.com/technology/2024/sep/15/data-center-gas-emissions-tech
[10] https://epochai.org/blog/can-ai-scaling-continue-through-2030
[11] https://towardsdatascience.com/the-carbon-footprint-of-chatgpt-66932314627d
[12] https://www.washington.edu/news/2023/07/27/how-much-energy-does-chatgpt-use/
[13] https://www.semianalysis.com/p/ai-datacenter-energy-dilemma-race

apunta que Europa estará atada de manos por la realidad geopolítica y las restricciones regulatorias estructurales sobre energía.

Soluciones

Si bien la IA generativa tiene un potencial increíble, con lo mencionado anteriormente también se hace evidente la necesidad de buscar métodos más eficientes y sostenibles para entrenar y ejecutar estos modelos. Algunas posibles soluciones incluyen diseñar mejores algoritmos, optimizar los modelos existentes, diseñar nuevas arquitecturas hardware más eficientes, o utilizar fuentes de energía renovables o nuclear.

Progreso algorítmico

Uno de los enfoques principales es mejorar la **eficiencia de los algoritmos** utilizados en la IA. Los avances en técnicas de entrenamiento, como la optimización del uso de datos y el ajuste más eficiente de hiperparámetros, pueden reducir el consumo energético. Por ejemplo, métodos como el **entrenamiento por lotes (batch training)** y el ajuste fino con menor cantidad de datos han demostrado ser efectivos para disminuir el número de cálculos requeridos (Eraíz-Fontanil & Schelle , 2022[14]). La innovación en algoritmos de aprendizaje profundo, como los **modelos de atención más eficientes**, también puede reducir los recursos computacionales necesarios.

Optimización de modelos (compresión)

Para hacer los modelos más eficientes existen varias técnicas de compresión.

- **Quantization**: Este método consiste en reducir la precisión numérica de los pesos del modelo, cambiando de precisión de 32 bits a 16 o incluso 8 bits, lo que disminuye la carga computacional y el consumo de energía.
- **Prunning**: Involucra eliminar conexiones y nodos innecesarios en la red neuronal, manteniendo solo las partes más relevantes. Esto resulta en un modelo más pequeño y eficiente, reduciendo la cantidad de operaciones requeridas.
- **Knowledge Distillation**: Permite transferir el conocimiento de un modelo grande (modelo “profesor”) a uno más pequeño (modelo “estudiante”), manteniendo un rendimiento similar con un menor costo computacional.
- **Tensor Networks**: Estas técnicas permiten representar grandes redes neuronales de manera más compacta, disminuyendo el número de parámetros y cálculos necesarios para la inferencia y entrenamiento.

[14] Eraíz-Fontanil, F., &Arce, A.(2022).Progreso algorítmico en modelos https://epochai.org/blog/algorithmic-progress-in-language-models

Nuevas arquitecturas

El desarrollo de hardware especializado ha sido clave para mejorar la eficiencia energética. Las **Unidades de Procesamiento Tensorial (TPUs)** de Google y las **GPUs** optimizadas para IA de NVIDIA son ejemplos de hardware diseñado específicamente para acelerar cálculos de redes neuronales con menor consumo de energía. También se están explorando arquitecturas novedosas como los **procesadores neuromórficos**, las **unidades de procesamiento fotónico,** o el uso de ordenadores cuánticos, que podrían ofrecer mejoras significativas en eficiencia.

Fuentes de energía

Para reducir la huella de carbono generada por los centros de datos que entrenan y ejecutan modelos de IA generativa, muchas empresas tecnológicas están recurriendo a **fuentes de energía renovable** como la solar, eólica, geotérmica y la hidroeléctrica. Los centros de datos de compañías como Google y Microsoft están adoptando estas fuentes para minimizar el impacto ambiental. Por ejemplo, Google ha logrado que algunos de sus centros de datos operen con energía 100% renovable al implementar acuerdos de compra de energía renovable (PPA, power purchase agreements) o construir infraestructuras en ubicaciones con acceso a fuentes sostenibles, como es el caso reciente de la inversión de AWS en España por valor de 15.000 millones de euros.

El uso de energía renovable ayuda a reducir las emisiones de CO_2 asociadas con el consumo energético de la IA. Sin embargo, también plantea desafíos. Las energías renovables, como la solar y la eólica, son **intermitentes** y dependen de las condiciones climáticas, lo que puede afectar la disponibilidad de energía. Por esta razón, las empresas a menudo deben invertir en sistemas de almacenamiento de energía, como baterías, para asegurar un suministro constante, lo cual puede aumentar los costos.

En los Estados Unidos, el crecimiento exponencial en el entrenamiento y explotación de la IA generativa también ha llevado a la industria tecnológica a buscar otras soluciones. Una de las tendencias emergentes es el uso de **energía nuclear** como fuente de energía para estos centros, en parte debido a su reciente clasificación como **energía limpia**. A pesar de los debates en torno a la energía nuclear, una de sus mayores ventajas es su **bajo impacto en las emisiones de carbono**. A lo largo de su ciclo de vida, la energía nuclear genera muy pocas emisiones de CO_2, lo que la posiciona en los esfuerzos para reducir la huella de carbono de los centros de datos.

Por otro lado, y a diferencia de fuentes de energía renovable como la solar o la eólica, que son intermitentes, la energía nuclear proporciona una fuente de electricidad **constante**, lo que resulta crucial para los centros de datos que requieren operar 24/7 con gran estabilidad.

Adicionalmente, en los últimos años han surgido diferentes iniciativas del gobierno federal de Estados Unidos, y por actores privados. Por ejemplo, como se está planteando la modernización y reactivación de plantas nucleares cerradas[15], aprovechando infraestructuras existentes. También han surgido nuevas tecnologías como los **reactores modulares pequeños (SMR, por sus siglas en inglés),** reactores diseñados para ser más eficientes, seguros y escalables que las plantas nucleares tradicionales. Varias empresas tecnológicas y energéticas en EE.UU. están invirtiendo en el desarrollo de los SMR como por ejemplo Rolls-Royce[16], lo que podría proporcionar una solución flexible y libre de CO2 para satisfacer las crecientes necesidades energéticas de la IA generativa.
Aunque la energía nuclear presenta muchas ventajas para los centros de datos, también plantea desafíos, como los **costos iniciales de construcción y de operación, la gestión de los residuos nucleares**, y las preocupaciones de seguridad por posibles accidentes, y por lo tanto la percepción pública negativa hacia esta fuente de energía.

Transparencia y sostenibilidad

Dada la creciente preocupación por el impacto energético y ambiental de la IA generativa ha surgido la necesidad de una mayor **transparencia** de reporte de estos modelos. La transparencia implica que las organizaciones no solo revelen el tamaño y las capacidades de sus modelos, sino también los detalles sobre su consumo energético y las emisiones de carbono asociadas al entrenamiento y uso de estos modelos.

Para abordar este asunto, se han propuesto diferentes iniciativas y herramientas enfocadas en medir y reportar el consumo energético de los modelos de IA, como por ejemplo Carbontracker[17]. Esta herramienta mide las emisiones de CO2 asociadas con el entrenamiento de modelos de aprendizaje profundo. Carbontracker ayuda a los investigadores y desarrolladores a monitorear el consumo energético durante las etapas

[15] https://www.xataka.com/energia/microsoft-reabrira-central-nuclear-que-lleva-cerrada-2019-quiere-para-alimentar-su-inteligencia-artificial
[16] https://www.rolls-royce.com/innovation/small-modular-reactors.aspx
[17] https://carbontracker.info/

de entrenamiento, facilitando la identificación de prácticas más eficientes y el cálculo del impacto ambiental del modelo.

Por otro lado, inspirándose en las etiquetas de eficiencia energética como las que se ven en electrodomésticos o bombillas, se ha propuesto una iniciativa denominada “AI Energy Stars”[18] para clasificar los modelos de IA según su eficiencia energética. Esta clasificación proporciona una métrica estándar que evalúa el consumo de energía en función de factores como la cantidad de horas de GPU necesarias para el entrenamiento, la energía consumida durante la operación y las emisiones de carbono generadas.

Un sistema de clasificación energética permitiría a los investigadores y empresas comparar diferentes modelos y tomar decisiones más conscientes sobre qué modelos desarrollar o utilizar. Al etiquetar los modelos de IA con una clasificación similar a las de las “Energy Stars” en productos electrónicos, los desarrolladores podrían ser incentivados a optimizar sus modelos, reduciendo su impacto ambiental y promoviendo el desarrollo de IA más sostenible.

A nivel nacional, la Secretaría de Estado de Inteligencia Artificial (SEDIA), con la ayuda de la Asociación Española de Normalización (UNE), han puesto en marcha el grupo de trabajo CTN-UNE 71/SC 42/GT1 "Evaluación de la eficiencia energética de los sistemas de inteligencia artificial" que tiene como objetivo definir las especificaciones UNE destinadas a medir, evaluar y mejorar la eficiencia energética de los sistemas de IA, centradas en las fases de entrenamiento e inferencia.

[18] https://www.nature.com/articles/d41586-024-02680-3